DOCUMENT RESUME

ED 169 066 TM 007 341

AUTHOR Bejar, Issac I.
TITLE Applications of Adaptive Testing in Measuring Achievement and Performance.
PUB DATE [76]
NOTE 8p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Complexity Level; *Confidence Testing; Factor Analysis; Scoring Formulas; Simulation; *Testing; Test Items; Test Reliability
IDENTIFIERS Computer Assisted Testing; *Latent Trait Models; *Tailored Testing

ABSTRACT

The concept of testing for partial knowledge is considered with the concept of tailored testing. Following the special usage of latent trait theory, the word valdity is used to mean the correlation of a test with the construct the test measures. The concept of a method factor in the test is also considered as a part of the validity. The possible effect of scoring for partial knowledge on such hypothetical tests is considered together with the logic of these hypotheses. The application of latent trait theory to a mathematical model is used to provide estimates of the expected gain in information as a function of the increase in inter-item correlations. Finally, these concepts are combined with the concepts of tailored testing. Two aspects of tailored testing are considered, tailoring test length and tailoring test difficulty. The possibilities of adapting tailored testing to non-dichotomous item scoring are considered in order to adapt tailored testing to the use of partial knowledge in the test score. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Isaac Bejar

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED169066

TM007 341

Applications of Adaptive Testing in Measuring Achievement and Performance

by
Isaac I. Bejar
University of Minnesota

Introduction

Achievement testing consists of locating individuals on an achievement scale. Usually, to interpret achievement test scores, a transformation is applied to the scores which allows an interpretation in terms of the relative standing of an individual with respect to the norming group. In many instructional settings, this interpretation is not adequate and, as a result, a demand for more concrete kinds of interpretation has emerged. The frequency with which criterion-referenced testing, mastery testing and similar approaches are used is evidence that the suggestion has been welcomed by test users.

What is unique about these testing procedures is that the items that constitute the test are sampled from a population of items which is isomorphic with the objectives of the instructional program on which we want to measure achievement (Shoemaker, 1975). Because of this, it is possible to interpret scores in terms of what the student can do in relation to the objectives of the instructional program.

Undoubtedly, this attention to content is bound to increase the quality of test scores. Today I'd like to describe our efforts at the University of Minnesota to improve achievement testing in general, including criterion-referenced testing approaches, by means of more refined response procedures as well as by adapting the test to the individual.

Background

Most psychometric theory assumes dichotomous scoring; that is, responses are classified as either correct or incorrect. However, knowledge is seldom binary, and by proceeding as if it were, partial knowledge is not given due recognition. If, in fact, partial information is present, then extracting it should lead to more valid and reliable scores.

The research literature, however, does not support the last statement. The results of the typical investigation show that while reliability is usually increased by taking partial knowledge into

---

This research is supported by Contract N00014-76-C-0627. NR 150-389, with the Personnel and Training Research Programs, Office of Naval Research, David J. Weiss, Principal Investigator.

account, the validity of the scores remains the same or even diminishes. Such findings are usually interpreted as evidence against the usefulness of the assessment of partial knowledge. To me, they indicate that something is amiss, for example, that the test and the criterion are not unidimensional.

To illustrate, consider two tests, A and B, measuring a single construct. Both A and B correlate .60 with the construct. This can be summarized as follows:

$$\Lambda = \begin{matrix} \text{Test} \\ \begin{bmatrix} .60 \\ .60 \end{bmatrix} \end{matrix} \begin{matrix} \\ \text{A} \\ \text{B} \end{matrix} \qquad [1]$$

Then the intertest correlation matrix can be expressed as in Equation 2,

$$\Sigma = \Lambda\Lambda' + \Psi, \qquad [2]$$

which in this case becomes Equation 3:

$$\Sigma = \begin{matrix} \Lambda \\ \begin{bmatrix} .60 \\ .60 \end{bmatrix} \end{matrix} \begin{matrix} \Lambda' \\ \begin{bmatrix} .60 & .60 \end{bmatrix} \end{matrix} + \begin{bmatrix} .64 & .00 \\ .00 & .64 \end{bmatrix}$$

$$= \begin{bmatrix} .36 & .36 \\ .36 & .36 \end{bmatrix} + \begin{bmatrix} .64 & .00 \\ .00 & .64 \end{bmatrix}$$

$$= \begin{bmatrix} 1.00 & .36 \\ .36 & 1.00 \end{bmatrix} . \qquad [3]$$

If we refer to the off-diagonals of $\Lambda\Lambda'$ as validities and to the diagonals as reliabilities, in this case both A and B have a reliability equal to .36 and validity of .36. Now suppose Test A is administered under conditions that allow for partial knowledge and that, as a result, its correlation with the construct goes from .60 to .70. Following the same procedure, we now find that the reliability of A is .49 while that of B remains at .36, and that the correlation (validity) has gone up from .36 to .42. In short, when there is a common factor between two measures, an increase in the reliability of one of them will lead to an increase in validity. This is not so when more than one factor is common.

To illustrate this, assume that Tests A and B, both administered conventionally, have in common, in addition to the construct, a method factor, and that both correlate .40 with it. That is,

$$\underset{2\times 2}{\Lambda} = \begin{matrix} \theta_c & \theta_m \\ \begin{bmatrix} .60 & .40 \\ .60 & .40 \end{bmatrix} & \begin{matrix} \text{Test} \\ \text{A} \\ \text{B} \end{matrix} \end{matrix} \qquad [4]$$

Assuming that the construct and the method factor are uncorrelated in the population, the correlation matrix for A and B, according to the model in Equation 2, is given by

$$\Sigma = \overset{\Lambda}{\begin{bmatrix} .60 & .40 \\ .60 & .40 \end{bmatrix}} \quad \overset{\Lambda'}{\begin{bmatrix} .60 & .60 \\ .40 & .40 \end{bmatrix}} + \overset{\Psi}{\begin{bmatrix} .48 & .00 \\ .00 & .48 \end{bmatrix}}$$

$$= \overset{\Lambda\Lambda'}{\begin{bmatrix} .52 & .52 \\ .52 & .52 \end{bmatrix}} + \overset{\Psi}{\begin{bmatrix} .48 & 0.00 \\ 0.00 & .40 \end{bmatrix}}$$

$$= \begin{bmatrix} 1.00 & .52 \\ .52 & 1.00 \end{bmatrix} . \qquad [5]$$

The validity is .52.

Now suppose that Test A above is again administered under conditions that allow for the scoring of partial information and that, as a result of this, its correlation with the construct becomes .70. At the same time the correlation of Test A with the method factor drops from .40 to .20; i.e., Λ becomes

$$\Lambda = \begin{bmatrix} .70 & .20 \\ .60 & .40 \end{bmatrix} \quad \begin{matrix} \text{Test A (with partial knowledge)} \\ \text{Test B} \end{matrix} \qquad [6]$$

and

$$\Sigma = \begin{bmatrix} .53 & .50 \\ .50 & .52 \end{bmatrix} . \qquad [7]$$

Thus, as a result of introducing partial knowledge, the validity was reduced from .52 to .50. However, it is clear that this seemingly disappointing result is not inconsistent with the true improvement that occurred, namely an increase of the correlation with the construct.

Although this example contains many assumptions, it seems that something similar occurs with real data. Hakstian and Kansup (1975) compared the validity of a verbal ability test administered under conventional and elimination scoring (Coombs, Millholland & Womer, 1956) instructions. Validity was defined as the correlation with school grades in language arts. This correlation was .49 under

conventional administration and .39 under elimination scoring. However, the correlation with another verbal ability test was .59 under conventional scoring and .67 under elimination scoring. Thus, defining validity as the correlation with school grades, elimination scoring appears to be less valid; but defined as the correlation with another verbal ability score, elimination scoring is more valid. These two findings are not contradictory but simply provide evidence of the fact that school grades and test scores are not unidimensional.

Advantages of Using Partial Information

In short, I think a critical review of the literature will convince most that the question is not whether partial knowledge scoring improves the validity and reliability of test scores but rather under what conditions are gains to be expected, and how large those gains are likely to be, in particular whether they are large enough to offset any increase in testing time. It stands to reason that if methods for the assessment of partial knowledge are to yield improved test scores, the tests must be such that there will be an opportunity for partial knowledge to emerge. With few exceptions, most notably Coombs, *et al.*, the presence of partial knowledge is never tested. Some theoretical results suggest that when partial knowledge is allowed to emerge and it is scored, dramatic improvements in test scores follow.

To illustrate this, I computed the information functions of two latent trait models. (You will recall that information at a given point on the underlying trait is the reciprocal of the variance of the maximum likelihood estimate at that point. Therefore the larger the information value, the more precise our estimate of the location of an individual on the trait.) One of the models uses the two-parameter normal ogive which is appropriate for dichotomous scoring. The other model was Samejima's (1969) graded response model, which is an extension of the two-parameter normal ogive to polychotomous scoring. You may think of the information of the graded model as the case when partial knowledge is taken into account, whereas the information provided by the dichotomous model is that provided when partial information is ignored.

To simplify the comparison, I computed for each model the mean information assuming that the underlying trait was normally distributed. The ratio of the mean information for the graded model over that of the dichotomous model for several levels of test homogeneity is seen in Table 1. For example, at r=.30 the ratio is 1.42. This means that, on the average, the use of partial knowledge will be 42% more informative than if it is ignored. Note that this improvement, due to incorporating partial information into the scores, increases as the discrimination of the test increases. In other words, the better the test, the more it will benefit from adding partial knowledge.

Table 1
Ratio of Mean Information of Graded to
Dichotomous Model, as a Function of Inter-Item Correlation

	Inter-item correlation					
	.30	.40	.50	.60	.70	.90
Ratio of mean information	1.42	1.43	1.48	1.52	1.58	1.90

The advantages derived from taking partial knowledge into account can only materialize under the proper conditions. In the conventional testing situation, even though partial knowledge influences which alternative is chosen, the response is scored as correct or incorrect. One way of allowing credit to be given for partial knowledge is to instruct testees to segregate alternatives into different categories. Coomb's procedure is an instance of this approach where the categories are "correct" and "incorrect". Other categories are possible, though, for example, verbal items may be classified as synonyms, antonyms, or neither.

Computerized Testing

Recording and scoring responses to this kind of item is not, however, convenient with paper and pencil administration. This brings me to another aspect of our research, namely the use of computers. One obvious use of computers is to handle the recording and scoring of responses, but as previous presentations in this symposium suggest, the computer can also be used to adapt or tailor the test to each individual.

These presentations, and indeed most of the research in computerized adaptive testing, are oriented toward ability measurement. In achievement testing, we should distinguish between two kinds of tailoring. One is tailoring the length of the test and the other is tailoring the difficulty of the test.

Tailoring the length of the test is appropriate in instructional settings where each individual is allowed as much time as necessary to complete a given unit of instruction. Under those conditions, individual differences with respect to knowledge are minimized and it becomes profitable to tailor the test in terms of length rather than difficulty. The research of Ferguson (1970) is an example of this type of tailoring. In his system, an individual is tested until he is classified into a non-mastery or mastery category. The statistical basis of this system is that of Wald's sequential likelihood ratio test. Ferguson's model assumes that the difficulty and discrimination of all items are the same. It is not known how sensitive the procedure is with respect to violation of these assumptions. Research addressed to this question is needed. It would also be desirable to study the possibility of relaxing the model to allow for unequal item difficulties and discriminations as well as allowing for polychotomous responses.

Although self-paced instruction has many advantages, limited resources often do not permit its full implementation. As a result, the sample under instruction will likely be heterogeneous with respect to achievement. Similarly, if we are testing for retention of achievement or for levels of achievement acquired prior to instruction, we will also find wide variation in performance. Under these conditions, tailoring the test to an individual's level of achievement will be more efficient than the conventional non-adaptive procedure, as the previous presentations suggest.

One of the major aims of our research is to combine the advantages of partial knowledge scoring and adaptive testing. Most of the research on adaptive testing at the University of Minnesota and elsewhere has been done in the context of dichotomous response models. The exceptions are to be found in the work of Bayroff & Anderson (1960), Wood (1971) and Samejima (1975).

Bayroff & Anderson seem to be the only ones to have actually implemented an adaptive testing strategy using non-dichotomous items. Essentially what they did was to branch an individual according to the correctness of the alternative chosen. Although they used a polychotomous item for the first item only, this can be readily extended to include all items. Other branching rules are possible. Wood (1971) suggested that the optimal branching rule will administer as the next item the most discriminating of those items with a midpoint of adjacent categories closer to the individual's current estimate of achievement. Samejima (1975) carried out a simulation on live data of a similar procedure which she referred to as tailoring the dichotomization of the item to the individual. She noted dramatic improvements by comparing the plot of scores based on a uniform dichotomization and tailored dichotomization against the scores based on the polychotomous responses.

Summary

To summarize, one part of our research is concerned with the joint implementation of two recent developments in test theory: adapting the test to the individual and simultaneously extracting more information from each response by recording partial knowledge. The question that remains is whether sets of items can be constructed such that they will allow partial knowledge to be utilized without unduly increasing testing time. By next year's meeting, I hope to have the answer to this and other related questions.

REFERENCES

Bayroff, A.G., Thomas, J.J., & Anderson, A.A. Construction of an experimental sequential item test. Research Memorandum 60-1, Personnel Research Branch, Department of the Army, January 1960.

Coombs, C.H., Millholland, J.E., & Womer, F.B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 17-37.

Ferguson, R.L. A model for computer-assisted criterion-referenced measurement. Education, 1970, 91, 25-31.

Hakstian, A.R., & Kansup, W. A comparison of several methods of assessing partial knowledge in multiple choice tests: II testing procedures. Journal of Educational Measurement, 1975, 12, 231-240.

Samejima, F. Graded response model of the latent trait theory and tailored testing. Proceedings of the First Conference on Computerized Adaptive Testing. United States Civil Service Commission, Bureau of Policies and Standards. 1976.

Samejima, F. Estimating latent ability using a response pattern of graded responses. Psychometrika, 1969, Monograph Supplement No. 17.

Shoemaker, D.M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.